DOCUMENT RESUME

ED 224 841 TM 830 040

AUTHOR McArthur, David L.; Hafner, Anne L.

TITLE Modifying Test Bias Through Targeted Instruction. Methodology Project.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 82
GRANT NIE-G-80-0112
NOTE 78p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Asian Americans; Black Students; Grade 5; Hispanic Americans; Instructional Innovation; Intermediate Grades; Low Achievement; *Minority Groups; Problem Solving; *Reading Comprehension; *Scores; *Test Bias; *Test Coaching; Test Items; *Test Wiseness
IDENTIFIERS California Test of Basic Skills

ABSTRACT

Systematic but unanticipated differences in patterns of responses to a test between two or more groups is generally taken as evidence of test bias. This study assessed whether a carefully-targeted instructional sequence could influence the effects of bias. A reading comprehension test with items previously identified as biased for certain groups was administered to two samples of minority children. This was followed a week later by two in-class sessions of the instructional intervention. Participants were then retested on the same instrument, first at the end of the same week in which they received the intervention, then again 4 weeks later. This pre/post/follow-up repeated-measures design allowed analysis, both statistical and graphic, of bias characteristics as they arose between groups at any given testing session and within groups across time. Results indicated that the test materials were generally very difficult. A few items improved significantly from pretest to posttest, although this improvement diminished somewhat across time. California Test of Basic Skill items which were targeted by the intervention showed a stronger degree of change for the treatment group than for the control group. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED224841

Deliverable - November 1982

METHODOLOGY PROJECT

MODIFYING TEST BIAS THROUGH TARGETED INSTRUCTION

David L. McArthur and Anne L. Hafner

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Grant Number
NIE-G-80-0112, P3

TM 830 044

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

CCL 2

TABLE OF CONTENTS

	Page
List of Tables and Figures	i
Abstract	ii
Introduction to Studies in Test Bias	1
Statement of the Problem	
Test Bias	5
Sources of Bias	7
Related Literature	
Test Bias Research	11
Modifying or Reducing Bias	1[illegible]
Purpose of the Study	24
Method	
Research Design	25
Instruments	27
Intervention	28
Procedures	30
Analytic Tools	31
Results	39
Summary	53
References	55
Tables	59
Appendices	70

TABLE AND FIGURES

Tables

1	Means and Standard Deviations for Age, Reading Level, Total Score, and Subtest Scores by Group and Ethnicity
2	S-P Analysis by Group by Occasion (Pre-Post)
3	Proportions of Wrong Responses above the P-Curve, for CTBS Items, by Group by Occasion, (Pre-Post)
4	Proportions of Wrong Responses above the P-Curve, for Bellagio Items, by Group by Occasion, (Pre-Post)
5	Non-Responses by Occasion (Pre-Post)
6	P's and S-P Cautions for Items (CTBS)
7	P's and S-P Cautions for Items (Bellagio)
8	Selected Ratios and Tests of Proportions, CTBS
9	Selected Ratios and Tests of Proportions, Bellagio
10	Proportions of Respondents with 011 or 001 Permutations
11	Intercorrelations of Eight Pre/Post/Follow up Permutations for Thirty Items

Figures

1	Score Permutations Ordered by Item Difficulty, CTBS (Pre-Post) Only
2	Score Permutations Ordered by Item Difficulty

ABSTRACT

Systematic but unanticipated differences in patterns of responses to a test between two or more groups is generally taken as evidence of test bias. This study assessed whether a carefully-targetted instructional sequence could influence the effects of bias. A reading comprehension test with items previously identified as biased for certain groups was administered to two samples of minority children. This was followed a week later by two in-class sessions of the instructional intervention. Participants were then retested on the same instrument, first at the end of the same week in which they received the intervention, then again four weeks later. This pre/post/follow-up repeated-measures design allowed analysis, both statistical and graphic, of bias characteristics as they arose between groups at any given testing session and within groups across time. Results indicated that the test materials were generally very difficult. A few items improved significantly from pretest to posttest, but follow up testing indicated some falling off of scores. CTBS items which were targeted by the intervention showed a stronger degree of change for the treatment group than for the control group. Hispanic treatment group subjects showed a slight advantage over their non-Hispanic counterparts, and over all members of the control group in terms of change from pretest to follow up.

MODIFYING TEST BIAS THROUGH TARGETED INSTRUCTION

David L. McArthur and Anne L. Hafner
Center for the Study of Evaluation, UCLA

Introduction to Studies in Test Bias

The widespread phenomenon of lower than expected test scores for bilingual and other minority students may not occur because their ability levels are lower but because (a) schools do not provide appropriate, equal instruction to all groups, and/or (b) the tests used to assess student abilities unfairly estimate their abilities. If the latter is true, it is generally assumed that item bias can be detected by a combination of statistical and linguistic/cultural methods. Once identified, biased items ought to be amenable to task-relevant instruction. In the present study, task-relevant instruction was used to train students in reading comprehension problem-solving skills to reduce the performance gap between majority and minority children.

CSE identified "bias" in assessment as a major determinant of differences in test scores. In the 1980-81 fiscal year, four major analyses were carried out to address the question of bias. The first

looked at classical test theory and scaling methods, along with a method from Japan (Sato's Student-Problem method) for the statistical analysis of item bias. The second analyzed selected aspects of item bias: (content, linguistic, cultural and social) in the CTBS, English and Spanish versions. The third analysis examined a data set which contained scores of both English and Spanish language versions of CTBS for the same set of students. The last analysis focused on ratings made by Hispanic and non-Hispanic raters who reviewed essays generated by Hispanic and non-Hispanic students.

In the item bias study, McArthur surveyed the professional literature and found that many indicators of bias exist. Some, however, are very complex and require a large number of items. (See McArthur, 1981, for further discussion.) McArthur turned to Sato's S-P method of analyzing test performance to look at discrepancies between actual and ideal response patterns. McArthur also used analysis of distractors, test of proportions of correct scores for masters, test of chance responding by masters and test of differential attractiveness of wrong answers.

McArthur's premise, that item bias can be detected by statistical analysis of persons x items matrices, was validated by the fact that from one-fifth to one-third of items in the CTBS (English version and Spanish-language version) showed strong evidence of bias. Such systematic patterns of bias in test items are most likely the result of complex interactions of group and individual factors with one another and with the tests. In this study, the CTBS Level C (first and second grades) and Level 2 (fourth and fifth grades) were

administered in English and Spanish to 1,259 students in California. Spanish-speaking groups scored lower in all subtests. Spanish language groups found the items more difficult at both levels and engaged in patterns closer to chance responding more often than did English speaking groups. Spanish speaking groups also had more items with popular distractors.

McArthur's findings were supported by Cabello's (1981) analysis of linguistic and cultural sources of bias for biased items and those not judged biased. Five categories were used as possible sources of influence on item content: (a) mistranslation; (b) cultural bias; (c) linguistic bias; (d) low frequency word bias; and (e) unfamiliar context bias. A great many of the items showed more than one statistical indicator of bias. Removing items from which three or more statistical indicators turned up positive gave adjusted scores which were more similar between groups (i.e., the Spanish-speaking group moved closer to their English language counterparts on three of four subtests). In effect, removing the items modified or reduced the "bias" in the test. A substantial difference remained between scores for the Passage Comprehension subtest at Level 2.

Cabello scrutinized items identified as biased to locate potential sources of bias, such as quality of the translation, curricular relevance and cultural interference. She found popular distractors attributable to mistranslation problems and cultural interference. Curricular relevance was not found to be a problem. Also, the types of tasks elicited by the test questions were examined.

Cabello found the most difficult types of questions were those requiring a student to infer the main idea, a character's feelings or the meaning of a metaphor. Next in difficulty were determination of sequences of events and derivation of word meaning from text. Finding explicitly stated information was shown to be the least difficult task. In sum, from one-fifth to one-third of the items on the CTBS were found to be biased . Removing items with the greatest number of statistical bias indicators helped the lower-scoring group move closer to the majority group scores. A large difference remained in reading comprehension in the higher grades (4th and 5th). More bias was found in the complex inferential items than in easier recognition and recall items.

If a test is incorrectly estimating a certain group's ability and that group's responses show systematic differences from responses of other groups, researchers can do several things to deal with this problem. They can remove items showing statistical and/or content indicators of bias and use only neutral items (assuming the remaining items consist of a sufficient number and range of difficulty and domains to satisfy test specifications). However, for this solution, a very large item pool would be needed. Researchers can also leave the "biased" items intact and attempt to limit the effects of bias. One way to do this is through targetted instructional sequences designed to teach the skills and objectives of the tests. This is the path CSE chose to take.

There is evidence in the literature that cognitive dimensions (verbal ability, strategy use or transfer) are associated more

strongly with test performance than are socioeconomic status or other demographic characteristics (Ulibarri, 1981). Differential performance may be a function of instructional background or different repertoires of cognitive skills and strategies.

CSE is now looking at ways of limiting effects of bias for different groups through the use of cognitive skills and instructional strategies. By providing children with the necessary reading comprehension strategies, CSE hopes to control for factors relevant to taking a test.

Under the assumption that teaching which directly addresses only the subject matter content (for example, all about abalone) is wasteful and may not be effective or carry over to other questions, this pilot study focuses on general reading comprehension problem solving skills that can be used across situations and across tests. In the past year of the present study, several different strands of theory were woven together to come up with a unique approach to test bias. In particular, CSE is moving beyond research on ways of identifying item bias to research on ways of modifying or reducing the effects of bias on items previously identified as biased for certain groups. It also seeks to ascertain whether bias is due to content or to item type.

Statement of the Problem

Test Bias. Systematic but unanticipated differences in patterns of responses to a test between two or more groups is generally taken as evidence of test bias. Test bias is a general measurement term which is used to refer to many things. For example, it can refer to

statistically-defined bias, sociocultural bias, linguistic bias, construct bias, predictive bias, or content bias. Definitions of test bias and item bias continue to proliferate in the literature. Yet, although bias is considered to be many "things", in fact it is not a thing but an abstract property or quality that is often used to explain test score differences. These differences can be identified either by statistical means or by face validity.

Because test score differences between groups persist, there has been a long-standing concern that tests are biased against certain groups. Test critics (such as the National Education Association and the various truth-in-testing groups) maintain that test scores reflect socioeconomic status, opportunity and education, not ability or aptitude. Test supporters (such as Jensen, 1980) say that research has shown these tests are valid for different groups and there is no large scale consistent statistical bias against minority groups. Of course, the two groups use different data and manipulate their data in different ways to arrive at their conclusions.

Three major schools of thought have evolved on the bias issue. First are those who take a strict statistical or psychometric approach. These people believe a test is unbiased when certain statistical characteristics of test data are invariant in the different groups to which a test is administered (see Jensen, 1980). In other words, a test is unbiased when items in a test behave alike in a statistical sense when administered to these groups.

A second group of theorists is content or face validity-oriented, and focuses on social and cultural concerns (see Williams, 1971).

Members of this group argue that a test is unbiased when it (or a sample of its items) does not discriminate between groups of respondents on the basis of cultural-specific knowledge. Differences between groups are the result, then, of variance in identification with cultural values or knowledge rather than variance in the mental ability being tested.

A third school of thought is concerned with group equity concerns or bias in test use. Some proponents press for tests tailored to specific populations, while others argue that bias in the practice of testing here denotes prejudice, and thus inequity.

The Center for the Study of Evaluation's work in test bias has adopted the view that none of these positions leads to a general solution of the bias problem. The divergent concerns should be merged. Along with statistical detection and analysis of bias, consideration of the cultural characteristics of the target population and the nature of its instructional history is deemed important. CSE has thus approached the question of test bias in both apsychometric and a content-analytic manner. CSE uses test bias to refer to "a systematic but unanticipated pattern of responses to a multiple-choice test found for an entire group of test-takers" (McArthur, 1981, p. 2). A different systematic pattern of responses implies there may be differences between two or more groups in abilities or skills or in cultural or linguistic issues related to the use of language in the test.

Sources of Bias. The first of three major sources of bias that have been identified is the language and thought patterns of test

makers. This source is often interpreted as context, as in the case wherein authors use words or expressions that are commonly used by members of their group. In addition, some test items may be put in a "context that is less familiar to one group than to another; in which case the items are likely to be more difficult for students in the former group, even though these aspects of the items are not part of what the test is meant to measure" (CAT, Tech. Bulletin, CTB/McGraw Hill, 1980, p. 140).

The language of test makers may lead to cultural interference. This refers to the idea that different aspects of a particular culture may interfere with a student's comprehension of, and performance on, an item. An example of this was found to be in favor of Spanish-speaking children by CSE. The item is drawn from the CTBS level C vocabulary test. It asks students for the synonym of "happy". The correct response is "gay". Students taking the Spanish test selected the correct response because gay in Spanish (feliz) does not connote "homosexual". Many of the English-speaking students, however, chose other responses because they were all too familiar with the colloquial meaning of gay. This item could be considered culturally biased against English speakers.

Cultural interference is almost inevitable in a test, as the test constructor assumes that all the children will perceive the same implied values from a passage. There may be nothing "culturally biased" about a passage, but a question referring to it may be biased because it necessitates knowledge external to the passage which may

vary from culture to culture (Cabello, 1981). If members of one cultural or ethnic group interpret a feature of a reading passage differently from others, they may choose a different answer than the one deemed to be "correct" by the developer. The possibility here arises that groups may be using different strategies to arrive at their answers.

The second major source of bias is the procedure used to try out items. Speaking for a test-maker, Green (1975) admits that one major source of bias is the procedure used to try out test material. Extensive tryouts eliminate some of the bias, but not all. Because of the nature of the item tryout and item analysis process, the characteristics of the majority of the tryout group will most likely be overwhelmingly represented. Any groups which are clearly under-represented will be shorted. The procedures used may result in a test which yields unfair scores for students differing from the majority in the tryout group. To deal with this problem, Arthur Jensen in his recent book, Bias in Mental Testing, recommends a standardization procedure that is sensitive to the differences displayed by various subgroups:

> Proper standardization for different subgroups should consist of comparable item selection procedures performed separately within each subgroup. The subgroups should be approximately equal in size, or at least each one should be large enough to permit comparable statistical inferences regarding the psychometric properties of the test. Only in the final norming (i.e., the computation of normalized standardized scores) for the composite sample should the subgroups be combined in proportion to their numbers in the general population (Jensen, 1980, p. 373).

Jensen's recommendation may be useful, but he gives no example of a test constructed in this way. It remains evident that the item selection process creates the very real danger of systematic bias.

The third major source of bias is the fact that test items may not be testing the same thing in different groups. For example, a reading comprehension test may be a measure of comprehension for Anglos but a measure of vocabulary for Hispanics. CTB/McGraw Hill, as one of the nation's major test publishers, believes that an achievement test is biased "when it systematically produces unfair scores for a particular group, a result that can occur only if the item measures something different for that group than for others" (Green, 1975, p. 36). A related problem is that the test may provide over- or under-estimates for a group's ability level. This is clear in the case of overprediction of Black and Chicano grades, and underprediction of women's grades.

We also need to ask which factors (outside of real differences in the ability tested) in students' histories lead to differential performance or to systematic bias? Primary factors that have been identified include background experiences, guessing, motivation, exposure to subject matter, coaching or teaching to the test, use of a particular strategy, differential course taking, or instructional history. Some of these are characteristic of a child's home background; others are determined by personality characteristics or individual differences and still others by classroom and school characteristics. Some of these are easier to influence than others.

For instance, background experiences in students' lives and motivational factors are hard to control.

Related Literature

Test bias research. A substantial literature has developed around the term "item bias" in the search for a single best all-purpose indicator which always reveals bias whenever systematic discrepancies in performance between groups are found. Many methods have been proposed and many studies conducted (cf. reviews in Berk, 1982; Subkoviak, Mack, & Ironson, 1981). Increasingly complex techniques have been set forth for the detection of bias in items (as previously discussed). No one approach has, however, explained why some items are biased and others are not.

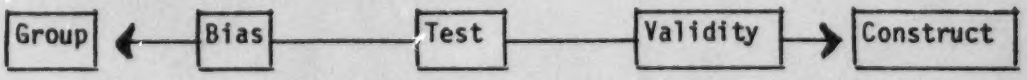
It may be helpful to review several definitions of bias for clarification. According to Petersen (1980), in an unbiased test, all the items would measure the same trait or ability and would be equally reliable and valid for all groups. It would also show orderly variation in the relative difficulties of the items, and be responded to in an orderly manner by every individual. One example of the outcome of this ideal is the familiar Guttman scale. Since most, if not all, tests are biased to some degree, the focus then turns to defining and identifying bias.

The Comprehensive Test of Basic Skills (CTBS) Manual puts forth its definition of bias. A test is biased "when it systematically produces unfair scores for a particular group, a result that occurs only if the test measures something somewhat different for that group

than for others" (CTBS Manual, p. 2). As noted previously, CSE considers bias to be "a systematic but unanticipated pattern of response to a multiple choice test found for an entire group of test takers". In this scheme, a biased item is one that functions differently for a subgroup of students.

What exactly does this mean? Here, we must draw from the concept of validity. Psychologist Nancy Cole (1981) maintains that questions of bias are basically questions of validity. NIE researcher Daniel Ulibarri, however, has clarified the distinctions between test bias and test validity (Ulibarri, 1979). He makes an analogy from the situation in the physical sciences to the one that exists in psychometrics.

As seen in the diagram above, in the physical sciences, test validity concerns whether an instrument is measuring the construct it purports to measure. Bias refers to the relationship between the instrument and the object of measurement (i.e., does a scale measure an object faithfully?). The relationship in psychometrics is illustrated in this diagram:



In psychometrics, test validity refers to the relationship between the test and the construct (i.e., does the test measure the construct?). Test bias refers to the relationship between the test and a group or groups. In other words, is the test providing over- or under-estimates or is it measuring the construct in different ways for two groups?

In related research, CSE researcher Beverly Cabello (1981) determined that cultural schemata (or context) present problems to second language learners such as Mexican-Americans. This brings up the possibility of bias in that the intended objective of the item (to test reading comprehension) would not be fulfilled, either because the item measures background knowledge or schemata or because the reader interprets schemata in a manner not intended by the developer. In this case, the item may be testing different things in two groups: in the majority group, it could be testing reading comprehension ability; in the minority group, it could be testing problem solving or schemata skills related to reading comprehension. Cabello goes on in her paper to present evidence of bias in items which can be attributed to infrequent vocabulary or content, translation problems, value-laden vocabulary, divergent interpretation of the same content or concepts with divergent referents.

Cabello explores further the concept of cultural interference in reading comprehension. She defines cultural bias as a phenomenon which occurs when there is a mismatch between the culture represented in a test item and the culture of the test taker. The possibility

that a test may be tapping students' knowledge of the test developer's culture as well as reading comprehension indicates a possible source of bias. All test items and passages contain semantic, syntactic, structural and content "cues" which can be helpful in solving the problem. Cultures vary greatly in the type of cognition and social processes which are valued, and in commonly held views on abstract ideas such as time, family, and science. Today, we do not have conclusive evidence on how cultures differ in cognition. But a relatively recent concept may be helpful in framing our inquiry.

The concept of a "schema" can be seen as the framework through which a person sees and comprehends an event, object, or concept (Bransford, Nitsch, & Franks, 1980). Schema is another word for context, and can encompass sociological, cultural, physical, nonverbal, and visual perception. Research has shown that cultural context and schema act as roadmaps to guide a reader's ability to comprehend a passage (Just & Carpenter, 1977). Bransford and Johnson (1973) found that if the frame of reference or context is unclear for a piece of writing, readers will either create their own reference point or fail to comprehend. If readers are provided with a reference point or context, comprehension is made easier. Context puts a frame around a passage--a framework which provides clues to associations and aids in classifying and shaping the message perceived. Rumelhart (1980) suggests that comprehension of stories depends on knowledge of problem-solving schemata and of different types of story plots or schemes, and that instruction on comprehending narrative discourse could include this skill as an objective.

Although there are discrepant explanations of why groups differ and why bias arises, there is a rough consensus among researchers that schemata, context or manner in which one approaches a problem are important determinants of whether a child gets a question right or not. For instance, anthropologist Gregory Bateson interprets context as learning to learn to receive and interpret signals (Bateson, 1972, p. 249).

Here, we can talk about taking a test as a learning experience. A child receives a signal or message from the reading passage and item stem. He/she classifies the signal (or context) based on prior knowledge or skills. Then, the child makes a response. Whether a child gets the answer right depends on several things:

1. Whether the child receives the exact message the item writer intended to send;
2. Whether the child classifies and interprets the message correctly;
3. Whether the child shapes the patterns or context into a meaningful whole and reads the relevant "cues" in the question;
4. Whether the strategies, knowledge or information the child possesses is appropriate for solving the problem;
5. Whether the child knows what the question wants him/her to do.

Essentially, these demands have little to do with whether the child has reading comprehension ability.

Another researcher, Scheuneman (1980) argues that the major sources of the differences found in test bias studies are: (1) flaws

in the item or test to which subgroup members are differentially sensitive; and (2) genuine differences between groups which might be the result of the cultural characteristics of the group and which might or might not reflect valid differences in the ability measured. Two ways in which bias in a test reduces the probability of a correct response from a person are uncertainty as to the type of task required by an item and the presence of cues which make the distractors unequally attractive to members of different groups.

Uncertainty, Scheuneman hypothesizes, "can result from failure to provide sufficient explanation concerning the nature of the task required to successfully complete the item or by introducing material into the item stem or options which serve to confuse the respondent" (Scheuneman, 1980, p. 4). Here, ability to guess what is required is part of what is being measured, along with whatever the test is intended to measure. Scheuneman gives an example of unclear directions in the Otis Lennon School Ability Test. Black 5th graders missed items requesting opposites more often than would be expected by their score on other verbal items. Closer inspection showed that Blacks were more apt than Whites to select the synonym, which suggests that the meaning of the original word was known but that some uncertainty existed about the word "opposite".

Another problem which should be considered is the presence of various "test-wiseness" cues which can help some students to eliminate incorrect options without actually knowing the correct answer. Sometimes, Scheuneman notes, this can be done without reading the stem. Some groups may be better at picking up and reading these cues.

In order to avoid a narrow view of the complexity of detecting bias in items, Scheuneman suggests reviewing groups of items and noting similarities and differences to detect patterns that may account for unexpected performance differences. Large numbers of items, however, are needed to do this. Scheuneman found that among the items with bias indicators in the School Language pretests for the Metropolitan Readiness Test pretest study, six out of seven were found to be testing grammatical usage of negatives. Since there are obvious differences in Black dialect, this adds support to the hypothesis that these items were measuring something different for Blacks and Whites.

Scheuneman recommends performing statistical bias analyses on items before reviewing items for general systematic differences. Here, items with unexpected differences are first identified and reviewers then look them over to detect what is wrong with them.

Scheuneman does not feel that items identified as biased should automatically be dropped, but that various courses of action should be explored. A procedure for reviewing items identified as biased is demonstrated by Scheuneman (1980).

1. Put "biased" items on cards. Record item statistics for different groups. Ask: what is the direction of the bias?

2. Sort items into broad categories. These are usually those outlined in the test specifications, or content categories. Tabulate numbers of biased and unbiased items in categories.

3. Work with items from one category at a time, and review "biased" items - try to detect item flaws or clues suggesting explanations for differences.

4. Verify hypotheses by checking against the set of unbiased

items (are there differences?). (Example of negative items and blacks.)

5. Consider what action might be taken to correct problems revealed by analyses.

Although Scheuneman infers that type of item (for example, literal recognition item or infer main idea item) is of primary concern in organizing and reviewing "biased" items, there is evidence that the content of items, not type of item, may be the determining factor in systematic differences.

For instance, in a CSE study, in the CTBS Reading Comprehension subtest at level 2 (5th grade), two passages accounted for more than 40 percent of items classed as biased (9 out of 21). One was a passage on abalone in the sea and the other was about threshing wheat on the farm. Looking only at an item's surface content in isolation did not give any useful information in these cases, as in some CTBS stories, main idea or sequence of events items were found to be biased, and in others not biased (McArthur, 1981).

Educational Testing Service admits that content of items rather than item types can be an important determinant of systematic differences. ETS also admits that (at least for the SAT test) is is possible to influence the magnitude of differences between the sexes by controlling the selection of item material (Donlon, Hicks, & Wallmark, 1980, p. 19). This finding by ETS brings up the possibility that test makers can also control the magnitude of significant differences between ethnic groups by controlling the selectin of item material (for instance, items testing grammatical use of negatives). To some degree, this is now being done for sexes in the development of

aptitude and achievement tests in the grade schools. The assumption here is that even though girls may do better on some types of items, boys do better on other types, so it is probably better to even out the items to make the test fair. This in fact could be done for ethnic groups, although public policy considerations could be prohibitive.

Modifying or Reducing Bias

Most of the recent test bias research has focused on fair test use and on developing techniques to detect bias. There are a few theorists, however, that have initiated research which attempts to modify or reduce bias.

A study of test bias done by Ulibarri (1982) hypothesizes that some students may perform poorly on tests because they may use inappropriate strategies and miss solutions to simple problems. In such cases a test is measuring a particular learned skill rather than a general ability like reading comprehension. Ulibarri's hypothesis is that some minority groups perform less well on tests because they do not use the strategy the test-maker had in mind. Although studies have shown that minority groups are equal to majority groups in information-processing capacity, Ulibarri feels that a "culture-loaded" item is one which calls for different information-processing strategies on the part of minority test takers. Ulibarri and colleagues trained a mixed group of children to select and use an appropriate strategy to tackle certain types of problems, hoping thereby to reduce the "bias". The researchers used items

already identified as "biased". After training, raw scores of Black and Hispanic children rose when retested. In addition, the childrens' teachers told the researchers they could see improvements in classroom performance. Ulibarri maintains that the results support the position that the training was affecting a learned skill and not the general ability the test was supposed to measure.

ETS researcher Janice Scheuneman developed a simple chi-square procedure for detecting and assessing bias in test items (Scheuneman, 1979). In this procedure, an unbiased item is considered "an item for which the probability of a correct response is the same for all persons of a given ability, regardless of their ethnic group membership" (Scheuneman, p. 145). Ability is the total score on a test or subgroup of items. The total score is divided into intervals. The percentage of people within each score range answering correctly is assumed to be an estimate of the probability of a correct response for those scoring within that range. A chi-square procedure is used to test the hypothesis that an item is unbiased.

Scheuneman applied her procedure in the 1976 revision of the Metropolitan Readiness Tests (MRT). She screened out of a large item pool those items with a high probability of bias. The subtests used measured aspects of visual discrimination, language proficiency and auditory discrimination. About 15 percent of the sample were Black. From a total of 555 items, 76 items were identified as biased. Within the set of biased items, 63 percent were from the language area, 16 percent were from the visual area, and 21 percent were from the auditory area.

After identification, the content of biased items was examined for reasons for bias, but causes were usually not apparent. One striking pattern appeared, however. Of 55 language items, 10 involved the use of negatives (i.e., "Mark the thing that is unopened."). Seven of these items were found to be biased and six involved the negative forms. The test authors were consulted, and the language objective was re-evaluated taking into account the fact that the use of negatives in Black dialects is known to differ from that in standard English. The test authors decided that for the first grade, the objective could be omitted and the test would still measure language skills. All ten items were dropped from the item pool, reducing the amount of bias for Black children.

It is not always possible, however, to delete items. Other educational researchers attempt instead to reduce the majority-minority gap through teaching students strategies.

Gerlen and Costar (1980) used a package program (Scoring High in Reading) to teach achievement test-taking skills in reading to 4th graders. Scoring High is a sequential reading program in which behaviors needed to score high on reading tests are taught. Examples of these skills include following group directions, considering every answer choice, using sound clues, eliminating inappropriate answer choices, identifying key words, and reasoning from facts. Each lesson focuses on one reading skill and gives extensive practice.

In this study, the control group followed the regular curriculum and the experimental group used Scoring High for two months. The

Metropolitan Achievement Test was used as the dependent variable. Differences were not found to be statistically different. Although the control group originally had a slightly higher mean score (3.85 to 3.72) at the second testing, the experimental group had closed the gap (4.11 to 4.18). Teachers in experimental groups reported that classroom atmosphere was more relaxed, and that less test anxiety was viewed.

Teaching children problem-solving techniques or strategies in a particular area, reading comprehension, has become the focus of a large body of research spearheaded by UCLA researcher, M.C. Wittrock. Wittrock (1982) sees reading with comprehension as a generative process. He has found that generating relations between text and one's own knowledge or experience contributes greatly to reading comprehension. This may be, he theorizes, because it takes effort to generate, because our own experience is involved or because it directs our attention. The main purpose of Wittrock's model of generative reading is to bring together the text and the reader's knowledge and experience to increase reading comprehension and decrease the gap between high and low ability groups.

Although Wittrock does not focus on test bias per se his methods speak directly to the issue of reducing bias in items between groups. Wittrock's research has shown significant results in improving reading capacity. In several experiments done by Wittrock and colleagues (Wittrock, Marks & Doctorow, 1975; Doctorow, Wittrock & Marks, 1978; Linden & Wittrock, 1981) reading comprehension among public school

students was improved by 25 to 100 percent with generative instructional strategies. These strategies include both teacher-given headings, titles, summaries, main ideas and learner-constructed headings, titles, summary sentences, main ideas, causes and effects.

In one study (Doctorow, Wittrock & Marks, 1978), 400 6th-graders read stories from reading materials. Some groups were given headings for the paragraphs. Other groups were asked to generate summary sentences for the paragraphs after they read them. Other groups were given paragraph headings and asked to generate summary sentences. It was found that the group given the generative instructions and paragraph headings doubled the comprehension and retention attained by the control groups.

Singer and Donlan (1982) taught 11th-graders problem-solving schema for comprehending stories. The experimenters theorized that if readers were given direct instruction asking questions about a story, they would acquire the ability to use knowledge about stories to focus on information in a story and improve their reading comprehension skills. The experimental group was given three weeks of special instruction two days a week. On the first 2 trials, no significant differences were found, but on later sessions, the experimental group scored significantly higher. The researchers concluded that instruction needs to continue over more than one story per skill and also needs multiple trials. They argue that this is not teaching to the test, as no significant differences were found on the first two trials. Effects do not accumulate quickly.

In another study, Dee-Lucas and DiVesta (1980) gave college students, or asked them to generate headings, related sentences or topic sentences as they read passages. Generation of topic sentences produced the greatest enhancement of learning. In other research (Linden & Wittrock, 1981; Arnold, 1981) children instructed to generate associations (summary sentences, pictures, main ideas) during reading showed greater comprehension than children not instructed to do so. In sum, this research points toward the value of the use of problem-solving strategies in improving reading.

The use of such reading comprehension strategies appears to be useful in improving reading comprehension and at the same time in reducing the majority-minority gap. It also points out that the difference between the groups may not be in amount of ability but in degree of strategy-knowledge relevant to getting test items right.

Purpose of the Study

The purpose of this study was to investigate whether carefully developed instructional sequences, designed to teach the skills and objectives of certain commonly used tests, would limit the effects of bias on items previously identified as biased for certain groups of students. The central hypothesis of the study was that direct instruction on intellectual skills related to reading comprehension would reduce the extent to which biasing elements contribute to explanations of test bias. A further hypothesis was that treatment effects would be statistically significant on post-testing, and would be statistically stable from post-testing to follow-up for all participants.

What were the assumptions of the study? First, it was assumed that the tests used (CTBS, Bellagio) measure reading comprehension skills and are valid and reliable for different groups involved in the study. Second, we assumed equal ability in reading comprehension in various groups. Third, we assumed that item bias can be detected by a combination of statistical and cultural methods. Last, we assumed that item bias so identified can be influenced by instruction. Task-relevant instruction attempted to train students in reading comprehension problem-solving skills to reduce the performance gap between majority and minority children.

Method

Research Design. The design for this study was a longitudinal treatment-outcome design involving three repeated measures--pre-training, post-training, and follow-up, utilizing a no-treatment control group. For evaluation of the relation of ethnicity to test bias, both groups were constituted with a mix of Hispanic, Asian, and Black students. For evaluation of additional factors known to contribute to problems of test bias, records were kept of gender, reading level, and training session attendance. The study design is expensive in terms of potential attrition, as it demands three testing occasions of all participants, and two additional sessions for training group participants. However, the design was chosen for its ability to show, under controlled conditions, the nature of any effects on test performance due to factors of training, time lapse, status variables, and possible interactions among factors. The

present study was conducted across eight intact 5th grade classrooms, drawn from two Los Angeles area school districts. The study was targetted at the 5th grade level for several reasons. In previous CSE analyses, many more items were found to be biased at the CTBS level 2 (5th and 6th grades) than for the lower grades. Reading comprehension tasks at the lower level tend to be far more literal in nature, requiring only skills in recognition and recall. Additionally, our statistical analyses require that we avoid a built-in ceiling effect, in which most students achieve correct scores on most items.

Classes were selected for participation on the basis of four criteria:

a) the class was predominantly of Hispanic origin (Part 1 participants) or Asian origin (Part 2 participants);

b) sufficient numbers of students in the class were capable of understanding spoken and written English;

c) the class was not scheduled to be tested with the California Tests of Basic Skills (CTBS) during the present academic semester;

d) the testing and intervention schedule was suitable for the teacher and not overly disruptive for the students.

We chose to use Hispanic children, as last year's Test Bias study identified items in the CTBS which were biased for Hispanics. Although Hispanic children have made significant gains in reading performance and these gains have exceeded those of students nationally in certain reading areas, improvement was greatest on literal comprehension. Hispanic students' performance still remains below the national average (National Assessment of Educational Progress, 1982).

The total number of participants available in these classrooms was 120. A certain amount of attrition over repeated test administrations was expected to deplete the initial subject pool; the data which follows concerns only those participants for whom records show attendance on all test and training sessions.

Instruments

The test materials for this study were selected from two sources following a search for materials which (a) were in the area of reading comprehension, (b) were designed for the upper primary level, and (c) if possible contained some evidence of item bias. The CTBS is a well-known instrument with a number of subscales; for this study 18 items of the Reading Comprehension subtest at Level 2, Form S, were selected (items 1-5, 20-25, 33-39). These items represent multiple choice responses from items which address three separate passages. These CTBS items were followed by another twelve multiple choice items drawn from the Bellagio Reading Comprehension Test, a CSE-developed instrument utilizing multiple choice responses from items addressing four separate narrative passages (see Appendix A). The Bellagio materials do not have any items for which bias characteristics have been assessed statistically, but each passage from the CTBS contains one or more items for which McArthur (1981) was able to point of evidence of possible bias when comparing English and Spanish-language versions.

Specifications were developed for the inferential comprehension domain. This included content limits, the distractor domain for

selected responses and item format, based on the single objective: "When presented with a reading comprehension passage (either narrative or expository) and questions about it, the student will be able to select the correct answer from four alternatives". (See Appendix B)

To avoid infringement of copyright, the CTBS test booklets and answer sheets were amended by blanking out or stapling shut unwanted sections and inserting the photocopied Bellagio passages and items. The answer sheet was amended by crossing out unwanted response blanks and designating twelve response lines of the "Reference Skills" section as the appropriate lines for responding to the Bellagio items. Total volume of material for the student was seven pages of text, each containing one passage and three to seven items, the original CTBS test cover, instructions to students, and directions for the Reading Comprehension subtest, and one standard response sheet. Pupils were instructed not to write in the booklets; they were checked and reused for posttest and follow up. Fresh answer sheets were provided to each student on every occasion.

Intervention

A two-session intervention specifically aimed at a generic issue in test-taking for reading comprehension--the inference of main ideas and relationships--was developed for this project, in consultation with experts in instructional design. Several factors were considered in designing the intervention for this study. We worked with reading comprehension tasks, since in previous analyses more differences among groups were found in this type of task than in others such as

vocabulary. Teachers polled in the Early Childhood Education Study reported that reading comprehension was the most important skill area. The area discriminates among students better than other tasks.

In particular, our intervention handled the interpretation of main ideas from a paragraph or passage. We felt that detailed training of the content dimension was beyond the scope of the hypotheses of this project. Therefore, our intervention aimed at teaching reading comprehension problem-solving skills to reach the objectives of the test. The CTBS Teachers Guide notes that the four types of skill measured by the CTBS are: (1) recognition and/or application of concepts and techniques; (2) translation; (3) interpretation; and (4) analysis. These are the process objectives. The first (recognition) is the lowest skill and requires only deriving literal meaning. The second (translation) implies rewording, rephrasing or translating from one language to another. This skill is still rather low-level. Interpretation involves comprehension of ideas and perceiving relationships. Items of this type can run from simple to difficult. This class includes the skill of drawing conclusions and identifying the main idea. The Teachers Guide notes that skills in theseprocesses are probably the most crucial of the lower-order skills and are necessary preliminary skills for higher-order verbal skills (analysis, synthesis). The ability to identify and comprehend major ideas in a passage and to understand their interrelationships is the skill discussed here. Thus, we hoped to teach two of the major objectives (skill areas) of the test: the

interpreting of main ideas through training on problem-solving strategies, and inferring relationships.

The intervention was geared to reading comprehension problem-solving techniques in inferring main ideas and relationships. Materials and strategies were drawn from the research in generative reading by Wittrock. These strategies include both teacher-given headings, titles, summaries, main ideas and learner-constructed headings, titles, summaries and main ideas. These strategies, and others, were used in the intervention (see Appendices C and D).

Procedures

The time schedule for this project allowed a pre/post/follow up design to take place within an eight-week period, utilizing English-dominant children from intact classrooms in Los Angeles. The test contained 30 items and testing time was 30 minutes. Thus, total testing time for pre, post, and follow up was about 90 minutes. The tests were administered by teachers or aides in the classroom (see Appendix E). The intervention took about 45 minutes and was administered twice for a total time of 90 minutes. Post-testing on the same materials was conducted within one week of the interventions, and follow-up testing was then conducted one month later.

In all cases, classrooms were maintained intact, but normal absences meant that some students were absent at any one testing or intervention session (or combination) and thus provided incomplete data.

Analytic Tools

To achieve a complete understanding of the nature of the responses generated in this study, analytic tools which address the "static" aspects of test performance for each group on each occasion, and the "dynamic" aspects of change through time across occasions are required. Two separate methods of analysis were used in the exploration of possible bias, and its modification, in the present study. The first was the Student-Problem (S-P) method, a system for analyzing patterns of test performance on one occasion. This tool provides summary data which addresses in particular the orderliness of fit between test items and respondents. Such orderliness, or lack thereof, can be compared on an item by item or respondent by respondent basis across repeated testing occasions. The second tool addresses the permutations of changing scores across repeated test administrations, that is, from wrong response to right response or right to wrong, as the same respondent encounters the same item a second time. This analysis of score permutations can be used to evaluate inter-group concordance, and is discussed further below.

The Student-Problem score table analysis has been developed over the last decade by a group of educational researchers in Japan (Sato, 1974, 1975, 1980, 1981a, 1981b; Sato and Kurata, 1977; Kurata and Sato, 1981; Sato, Takeya, Kurata, Morimoto and Chimura, 1981). While the mathematics associated with derivative indices in this system are relatively complex, the S-P system itself is predicated on a simple reconfiguring of test scores. Rather similar analyses of student

performance on educational tests can be found in the professional literature of a half-century ago, but recent developments by Sato and colleagues represent significant improvements in both concept and execution.

Test scores are placed in a matrix in which rows represent individual respondents' responses to a set of items, and columns represent the responses given by a group of respondents to a set of items. The usual (and most convenient) entries in this matrix are zeros for wrong answers and ones for correct answers. Total correct scores are calculated for each respondent, and total number of correct responses are tallied for each item. Rows are reordered by descending total number of correct responses; columns are reordered by ascending order of difficulty of items. The resulting matrix has several aspects which are particularly convenient for a detailed appraisal of respondents or items, singly or collectively.

Two cumulative ogives are drawn over the matrix to form the framework for further analysis. Because the data is discrete, the ogives take on a stair-step appearance, but both can be thought of as approximations to curves which describe in summary form two distinct patterns embedded in the data. The first is a curve reflecting respondents' performance as shown by their total scores; the second is a similarly overlaid ogive curve reflecting item difficulties. In one special circumstance, the two curves describe only one pattern: if the matrix of items and respondents is perfectly matched in the sense of a Guttman scale, both of the curves overlap exactly. All of the

correct responses would be to the upper left while all of the incorrect responses would be to the lower right. However, as the occurrance of either unanticipated errors by respondents with high scores or unanticipated successes by respondents with low scores increases, or as the pattern of responses becomes increasingly random, the respondent or student curve (S-curve) and the item or problem curve (P-curve) become increasingly discrepant.

Sato has developed an index which evaluates the degree of discrepancy or lack of conformation between the S- and P-curves. This index will be zero in the special case of perfectly ordered sets, and will approach 1.0 for the case of totally random data.

The index, called the "disparity coefficient," is explained as follows:

$$D^* = \frac{A(I,J,p)}{A_B(I,J,p)}$$

where the numerator is the area between the S curve and the P curve in the given S-P chart for a group of I students who took J-problem test and got an average problem-passing rate p, and $A_B(I,J,p)$ is the area between the two curves as modeled by cumulative binomial distributions with parameters I,J, and p, respectively (Sato, 1980, p. 15; indices rewritten for consistency with notation of Harnisch & Linn).

The denominator is a function which expresses a truly random pattern of responses for a test with a given number of subjects, given number of items, and given average passing rate, while the numerator reflects the obtained pattern for that test. As the value of this

ratio approaches 1.0, it portrays an increasingly random pattern of responses. For the perfect Guttman scale, the numerator will be 0 and thus D* will be 0. The computation of D* is functionally derived from a model of random responses, but its exact mathematical properties have not been investigated thoroughly.

For any respondent, or for any item, taken individually, the pattern of scores reflects that row or column in relation to the pattern established by the configuration of sorted rows and columns. For any given individual respondent or single item, the response pattern may be "perfectly ordered" in the sense used above. The row or column shares a symmetry with the associated row or column marginal. An index of this symmetry which is stable across differing proportions is Sato's Caution Index C, which gives a value of 0 in the condition of "perfect symmetry" between row or column and row marginal or column marginal. As unanticipated successes or failures increase and "symmetry" declines, the index increases (a modification of the Caution Index, called C*, has an upper bound of 1.0). Thus a very high index value is associated with a respondent or item for which the pattern of obtained responses is very discrepant from the overall pattern established by all members of the set.

Harnish and Linn (1982) present the modified Caution Index as follows:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij}\, n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J+1-n_{i.}}^{} n_{.j}}$$

where $i = 1,2,\ldots,I$ indexes the examinee,

$j = 1,2,\ldots,J$ indexes the item,

$u_{ij} = 1$ if the respondent i answers item j incorrectly,

0 if the respondent i answers item j incorrectly,

$n_{i.}$ = total correct for the i^{th} respondent, and

n_{ij} = total number of correct responses to the j^{th} item.

Harnisch and Linn explain that the name of the index comes from the notion that a large value is associated with respondents that have unusual response patterns. It suggests that some caution may be needed in interpreting a total correct score for these individuals.

An unusual response pattern may result from guessing, carelessness, high anxiety, an unusual instructional history or other experiential set, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions.

A large value may also suggest that some individuals have acquired skills in an order which is not characteristic of the whole group. The index says nothing about the most able respondents with perfect total scores, because the "symmetry" condition is met. More importantly, if a respondent gets no item correct whatsoever, both the total score and the caution index will be zero since, again, the "symmetry" condition is met; in this situation the available information about the respondent is insufficient to make any useful diagnosis. Most persons, though, will achieve total scores between the extremes and for them the caution index provides information that is not contained in the total score. A large value of the caution index raises doubts about the validity of the usual interpretation of the total score for an individual.

The second primary analytic tool for the investigation of bias addresses the nature of the respondent's performance across repeated trials. It is based on tallying correct and incorrect responses to each item in terms of score permutations (cf., van der Linden, 1981). For evaluation of pre and posttesting only, four permutations are possible for any item: always correct (11), always wrong (00), moving from wrong to correct (01), and moving from correct to wrong (10). When three testing occasions are involved, eight possible permutations

of correct and incorrect scores can be achieved for a given item by a given student. These permutations, and the symbols used for the remainder of this paper are:

pre	post	follow up	symbol
correct	correct	correct	111
wrong	correct	correct	011
wrong	wrong	correct	001
wrong	correct	wrong	010
correct	wrong	correct	101
correct	correct	wrong	110
correct	wrong	wrong	100
wrong	wrong	wrong	000

The first and last of these permutations contain those respondents for whom the repeated trials have no ostensible effect; either the answer were consistently right (111) or consistently wrong (000). The second and third permutations contain those respondents whom intervention between pre- and posttesting might have made an impact, chance responding aside. The remaining permutations tally those respondents whose performance on an item across time is not readily interpretable with reference to the specific intervention, but more likely reflects partial knowledge, inconsistent task processing, guessing and/or chance. The proportions of those in each category can be contrasted for items across trials; it is also meaningful to assess certain ratios, such as the ratio of those moving from an incorrect response to a correct response vs. those moving in the reverse direction, from a correct response to an incorrect response.

Recent developments in nonparametric statistics are being examined for their applicability to this design. Kraemer's (1981)

procedure for calculating intergroup concordance conceptually is directly on target, but requires data in rank form where ranks are mostly unique, not merely binary as in the present study. The technical contributions of Mielke and colleagues (cf. Mielke, Berry, Brockwell and Williams, 1981) also require further study. Additionally, the treatment and control groups' performance at each separate test occasion was examined using Boolean Factor analysis, a technique based on Boolean manipulations of binary data. This procedure to date has not yielded the anticipated benefits but consultations with M.R. Mickey of UCLA's School of Medicine indicte that further iterations may bear fruit (Dr. Mickey originated BMDP8M, the only currently available Boolean factor procedure).

RESULTS

One hundred and twenty 5th grade students served as subjects in this study. The participants were drawn from seven classrooms across four separate schools. Their average age at the time of initial contact was 10.0, s.d = 0.57. Hispanics represented 68.6 percent of the sample, blacks 8.6 percent; the remainder was predominantly of Asian heritage. The number of participants who received the two training sessions and provided complete data on pre and post testing was 66. The total number of children who received both training sessions and provided data on all three testing occasions was 58.

The mean age, reading level, test and subtest total scores for the complete sample of respondents receiving pre- and post-testing is shown in Table 1. There are no statistically significant ($p<.05$) differences at this gross level of test performance. (Although it is evident that the two groups, even though schools were selected for ethnic comparability and classes were randomly assigned to treatment condition, may not be entirely comparable.) The bold-face values on Table 1 show the percentage of improvement from pre- to post-testing for each subgroup. For the CTBS subtest, the treatment group improves noticeably more than the control group; the Bellagio subtest reflects a mixed picture of change.

The next four tables address data from the 93 student who provided complete pre- and post-testing. Table 2 summarizes results

from the S-P analyses conducted on the treatment and control groups without reference to ethnicity. In both groups, the overall results point to a great deal of random guessing by all respondents on both occasions. The D* values, which index overall lack of comparability to a perfect Guttman scale, start at a high value and move even higher on post-testing. The average C_i* values for items, which index the degree to which patterns of responses are inconsistent across items, moves slightly up from a moderately high value on pre-test to post-test. The average C_i* values for persons, which index the degree to which particular respondents are unlike their counterparts, are also moderately high on both pre- and post-testing. The number of items and persons for which the associated C_i* or C_j* values exceed .300 is alarmingly high for both groups on both testing occasions. The items on this test ranged from moderately to very difficult for a majority of participants; the proportions of total scores which were no better than chance were 21.5% on pre-test and still 4.3% on post-test. The comparison of item p-values across repeared testings is contaminated by the degree of random responding. Instead, a comparison was used which selected those persons who, by evidence of their total scores and the configurations of items within the group, should have achieved a correct response on a given item. That is, turning to the S-P chart, each item's pattern of responses <u>above</u> the p-curve can be tallied. If that item is perfectly behaved in the Guttman sense, that response vector should contain nothing but correct responses. The proportion of wrong responses in this selected vector

changes from item to item as the p-curve changes, and as the item is more or less well-behaved. Table 3 shows, for pre- and post-testing, the proportion of wrong responses above the p-curve, for the CTBS items, by group by occasion. It offers a detailed picture of item behavior over repeated occasions: mean change for the two groups is nonsignificantly different but some items behave substantially different by group. Note that the target tasks of the treatment intervention are represented by half the items, and that two thirds of the items by previous analysis are suspected of having possible bias. Table 4 shows the same figures for the Bellagio items; the target tasks are represented by a little over half of the items, although no prior evidence is available to suspect bias.

Of the nine CTBS items which represent the target of instruction, five showed some improvement from pre- to post-test among the treatment group. However, two of these items showed equal or greater improvement between pre- and post-testing for the control group. The amount of improvement was not related to whether the item was one selected for possible bias, and was somewhat related to the item's position in the testing sequence (later items in both CTBS and Bellagio subtests generally have poorer p-values and more inconsistent response patterns).

Of the seven Bellagio items which represent the target of instruction, only two showed improvement from pre- to post-test in the treatment group, only one in the control group. Indeed, both groups show a very severe difficulty with several Bellagio items (in part

because of the missing response problem addressed below). Again, no ostensible relationship can be seen between nature of change or presence or absence of and nature of intervention.

One important ingredient in the nature of change from pre- to post-testing is the number of items for which a respondent fails to provide a response. Despite allotting time for testing appropriate to most 5th grade students, some participants in this study worked substantially slower and had trouble finishing all items. The average number of missing responses on pre-test was 4.10, the bulk of which occurred in the concluding items of the Bellagio subtest (see Table 5). The amount of change from pre- to post-test ranged from omission of four additional items to inclusion of 24 items omitted on the pre-test. (The latter occurred for an individual whose performance on both occasions was poor; on post-test he was observed to mark up most of his answer sheet more or less at random. As a consequence, his C* values are high on both pre- and post-testing S-P charts, indicating an anomalous performance.) Due to a variety of factors, such as familiarity with the test or the testing situation, the number of non-responses decreased substantially overall from pretest to posttest. Those whose performance on posttest was correct for a given item would be included in the 01 permutation even though they gave no response on pretest, thereby contributing to the appearance on the figures of a treatment effect.

A visualization of score permutations for the treatment group between pre- and post-testing is presented in Figure 1, which, for

purposes of clarity, portrays cumulatively the permutations for the CTBS items only, when items are sorted by increasing levels of difficulty. This test determination was made by ranking items by the number of respondents who succeeded in giving a correct response on both test occasions. This 11 permutation is shown as the initial white band on Figure 1. The opposite permutation, 00, is shown by the crosshatching at the bottom of the graph. The white band above, 10, represents the proportion of treatment group respondents who moved from a correct response on pre-test to an incorrect response on post-test--a move, of course, in opposition to the intent of this study. This proportion is remarkably constant regardless of item difficulty, and this most likely can be taken as an index of the amount of guessing.

The remaining band, shown by both stippling and horizontal lines, is the primary target of interest in the present analysis. The stippling shows the proportion of respondents in the treatment group who moved from a wrong response to a correct response (01) and the horizontal lines show the proportion of respondents in the control group in the same 01 permutation. For purposes of visualization, the latter has been centered on the former, so the degree to which stippling shows itself without the overlay is the degree to which the beneficial effect of treatment was achieved. This effect, unfortunately, is seldom large, and appears to be poorly related, at best, to the focus of instruction. The eight items most affected include only three of the nine items for which instruction was geared,

and is unrelated to whether or not the item was especially difficult or relatively easy.

At this time, we turn to the smaller sample of respondents who were involved in both treatment sessions and provided complete data for all three testing occasions. The three left-hand columns of Table 1 show the p-values for each testing occasion for the 18 items drawn from the CTBS. This was a difficult test for these children: two items (#16, #17) never rise above the level of chance responding across all three testing occasions. The overall average correct scores increase 10 percentage points from pretest to posttest, and the number of items at the chance level drops from nine to five. Fifteen of the eighteen items show an increase in average p-value from pretest to posttest, with one increase as large as 30 percentage points, although only a few of these changes actually represent statistically significant improvement, as will be examined shortly. Ten of those fifteen items with an improvement from pre to post show a decrement on follow up testing; however, the number of items at the chance level is reduced to four on follow up. The fourth column of Table 1 notes the twelve CTBS items for which prior evidence demonstrated one or more possible sources of bias. In terms of this minority sample, no significant distinction can be found between "biased" and "unbiased" items in either item p-values or in amount of increase or decrease in p-values from pre to post or from post to follow up.

The three left-hand columns of Table 7 show the average p-values for the twelve items of the Bellagio reading test. These values are

quite similar in both tests. However, the amount of increase from pre-testing to post-testing is generally smaller than that of the CTBS. Fewer items are at the chance level of responding, but every item for which there was an increase in p-values from pre to post showed a decrease from post to follow up. Three items (#8, #11, #12) never elicit more than a chance level of correct responses on any occasion. Unlike the CTBS, no previous evidence is available to index suspected bias in these items.

The S-P analyses for the CTBS and Bellagio items add another dimension to these results, specifically focused on the pattern of responding within each testing occasion. Overall, the D* values are quite high; they increase across testing sessions, suggesting many anomolous patterns of responding in all three testing sessions. In the three right-hand columns of Table 1 the C* caution indices associated with the CTBS items are given for each testing occasion. Note that the calculations are executed independently for each testing session, but that most of the items are relatively stable as to their C* values across occasions. It is important, however, to note that while an item can show no change in average p-value from one occasion to another, the C* index is free to vary. An example is CTBS item #4, initially correct only at the chance level, then correct at both post and follow up testing at the level .40. The item's caution indices show that correct responses occur in a highly anomalous pattern on follow up.

An item with very good stability in its caution indices is CTBS #6, which showed a modest increase in p-value from pre- to posttesting, followed by a trivial decline on follow up. These figures illustrate that the task of interpreting change in p-values across occasions is more complex because the particular pattern of correct and incorrect responses for each item, is not totally independent of the item's p-value, and represents a dimension of information about the item which is not available in average p-values alone. In essence, items with higher caution indices show that the respondents are behaving more erratically with regard to that item, and that correct (or incorrect) responses are being given by participants whose overall correct score would have suggested an incorrect (or correct) response to that item. For CTBS item #6, the pattern of responses as indexed by the C* values on pre, post and follow up, are consistently low. In contrast, CTBS item #1, which has about the same range of p-values, is flagged with higher C* values, which suggests that the respondents, while achieving about the same overall rate of success, did so in somewhat less predictable patterns.

For the Bellagio reading test, the D* values are also quite high, and increase slightly across occasions. The three right hand columns of Table 2 give the corresponding C* caution indices for the 12 Bellagio items and show several items for which the patterns of responding are fairly anomalous.

Now we turn to the analysis of score permutations across the three repeated testing occasions. Figure 2 is an amplification of

Figure 1, including more items and an additional testing occasion, thus more permutations, but is the same conceptual layout. Figure 2 portrays cumulatively the permutations for the CTBS and Bellagio items combined, when items are sorted by increasing levels of difficulty, as determined by decreasing numbers of respondents who achieved correct scores on all three testing occasions (111), shown at the top of the graph. The stipled section, immediately below, represents the number of respondents who moved from a wrong response on pretest to a correct response on posttest and were able to give the correct response on follow up as well (011). This permutation of responding is the one most suggestive of an effect due to the training program although chance responses undoubtedly contribute. The graph makes clear that the 011 permutation is minimal for most of the hardest items, and occurs with some strength only infrequently for the easy and moderate items.

The white band below the stipling represents the 001 permutation, achieving a correct score only on follow up. This particular response permutation occurs at a moderate rate for all but the easiest items, and is the result either of a delayed learning of the principles imparted in the training sessions, or, more likely, a degree of chance correct responding on the third testing occasion by a small but steady number of individuals. The participants tallied by the 011 and 001 permutations (shown by the stipled band and the white bank below) are the ony ones in this sample who can be thought to have shown any ostensible lasting effect due to treatment.

The band shown by waves represents that group of participants who got an item wrong on pretest, right on posttest, then wrong on follow up (010). That is, this band shows the number of respondents who may have achieved some learning in the course of the training sessions, but who did not retain those skills, or a memory for their earlier response, on follow up testing.

A surprisingly consistent percentage of the responses fall into this permutation. Moreover, it must be noted that only a minority of pretesting and follow up testing incorrect choices are the same choice: that is, most respondents showing the 010 permutation did not go back to repeat their original error after having selected the correct response once. (The "P.T. Barnum effect" in psychological research is a direct analogue, although with the opposite conclusion. Frequently, respondents in a repeated measures design have been reported to return to their original erroneous responses on follow up even when given unambiguous information about an item.)

On Figure 2, the white band below the waves, the vertically-lined band, and the white band below that, all represent anomalous response permutations. The first (101) indicates that a respondent appears to have selected the correct response without training, then loses it immediately after training, only to regain it four weeks later. The second (110) and third (100) suggest that training had something of a deleterious effect, resulting in delayed or immediate loss of the correct choice for an item. Far more likely, given the evidence from the D* values in the S-P analyses presented earlier, is that all three

of these permutaions represent chance correct responding. The percentage of respondents with one of these three permutations is small for every item, but taken cumulatively the three permutations account for at least the same proportion of responses as the more interpretable 011 permutation. In other words, the samples tended to decrease in their ability to give a correct response about as often as they improved it.

The final band on Figure 2 (shown by crosshatching) represents those respondents for whom one can say unequivocably that the training sessions had no effect. The permutation of 000 instead indicates choice of the same erroneous response three times over, or a choice which varies between one and another of the wrong responses. On average, persistence with the same erroneous response across all three testing sessions, expressed as a proportion to the number of 000 responses, occurred only 21 percent of the time. That is, very few respondents overall stayed with the same wrong choice across all three testing occasions. This figure, however, for the six most difficult items actually represents the same number of respondents who were able to get the item correct on every trial (111).

There are several ratios between permutations which are instructive. Table 8 presents for each CTBS item the ratio of the number of respondents who moved from a wrong response to a correct response (permutations 011 and 001) to the number who moved from a correct response to a wrong response (permutations 110 and 100). When this value exceeds 1.0, the item is one for which the "gain over time"

exceeds the "loss over time" effect: 13 out of 18 items demonstrate such a value. For the remaining items, the ratio is evenly balanced at 1.0, or weighted towards "loss". The relationship is close, but not exact, between these values and the amount of increment or decrement across the p-values for Table 6, since this computation specifically excludes both the 111 and 000 permutations for which the issue of "gain" or "loss" is moot, and the 010 and 101 permutation for which the interpretation of "gain" or "loss" depends on which time frame is taken as evidence. The second column on Table 8 shows the ratio of "always correct" (111) to "always wrong" (000), which is a summary term reflecting item difficulty for those unaffected by the training. When this value exceeds 1.0, the item is easier to always get right than always get wrong. The largest such values greater than 1.0 occur among the easiest test items but it should be noted that even a relatively easy item can be one which eludes a sizeable number of respondents on every testing occasion. Taken together, the two left columns of Table 8 address the impact of training in relation to the general tractability of the item. A large first term with a large second term reflect items which were not too difficult to master; a large first term with a small second term reflect items which generally could not be mastered with training.

Table 8 also shows the results of two statistical tests of the significance of the difference between correlated proportions. For this the permutations of responses were recast as dyads (11, 01, 10, and 00), one set representing the pretesting and posttesting responses

only, the other for posttesting and follow up testing only. Three positive changes prove statistically significant; all three are associated with a large ratio of "gain" over "loss" and a small ratio of "always right" to "always wrong". While there are some decrements in these z-values, none of the decrements proved statistically significant within the pre- and posttesting combination. On follow up, there are no significant additional improvements, and one item (#17) shows a significant decline. That item, it should be noted, was one of several which never rose above the chance level of responding, so statistical significance in this instance is not so readily interpretable.

Table 9 shows the corresponding "moving to correct" vs. "moving to wrong" ratios, "always correct" vs. "always wrong" ratios, and tests of correlated proportions for the 12 Bellagio items. Only five of twelve items show any "gain over time", and only one of the items shows a statistically significant change from pre to post (#2, in the negative direction). None of the changes from post to follow up are statistically significant.

Table 10 presents the proportions of respondents in the treatment and control group, respectively, who succeeded in moving from an initial wrong response to a correct response by follow up. The move is symbolized both by permutation 011 and permutation 001. Either permutation reflects the impact of taining and the passage of time for the treatment group, and the passage of time only for the nontreatment control group. Thus, any significant differences between proportions

should be able to be ascribed to an impact of training. Both groups experience a high degree of variability in these permutations, and the control group has several items for which no member within that group produced a move to the right answer on post-testing and held to that answer on follow up. The only significant difference arises in comparing the targeted CTBS items across groups ($t=2.62$, $p<.05$); differences between groups involving nontargeted CTBS items and all comparisons of Bellagio items were nonsignificant.

Is there a relationship across all thirty items between the various permutations of correct and incorrect scores? Table 11 shows the product-moment intercorrelations between the number of respondents sharing each of the eight pre-, post- and follow up testing permutations across the thirty items. The numbers in bold type represent polar opposites: as might be expected from inspection of Figure 2, there is a sizeable negative relationship between the 111 permutation and the opposite, 000. The remaining values are generally trivial. The values on this table suggest that there is only one meaningful relationship among the bands of performance activity shown in Figure 2--the number of respondents in the top band of "always correct" responses is inversely related to the number of respondents giving "always wrong" responses. Otherwise, these bands generally are not linearly related.

Certain problems with the current study must be addressed. The first is that, despite the volume of testing conducted and the attempts made to schedule as carefully as possible, normal school

absences caused a not unexpected attrition in total session attendance. This attrition affects the precision of all statistical decisions, and by the third testing session, has swallowed up unknowable amounts of performance variation. Additionally, data from those participants who missed even one testing or intervention session must be handled differently from those who missed none.

Summary

The results from the set of analyses indicate that the CTBS-Bellagio test materials were generally very difficult for this group. However, it is important to recognize that this group of 120 students may not be representative of fifth grade students in general, because the four schools involved were in the lowest quartile of scores on the 1982 California Achievement Program. (A larger sample, chosen from schools with a wider range of CAP scores, may substantively affect the nature of this pre-post-follow up data.) After two sessions of a targeted intervention in the classroom, there are a few significantly improved items when comparing posttest to pretest. In particular, CTBS items which were targeted by the focus of the instructional intervention, showed a stronger degree of change from pretest to post-test and follow up test for the treatment group than for the group which received no treatment. Follow up testing, however, tends to indicate some falling off of scores across both CTBS and Bellagio items, for both treatment and control samples. Moreover, the number of anomalous response patterns, when expressed either as an elevated caution index for students at each test session or as a

confounded response permutation (such as 101) across sessions, is not trivial.

Contributing factors to the anomalous response patterns and the general paucity of results include possible nonequivalence between subtests, small sample size, large degree of guessing, appreciable number of nonresponses, not enough testing time for some students, and the relative brevity of the treatment program. The two subtests may have been insufficient for the task of discriminating a training effect. The training progam may not have left a long-lasting instructional impact for many respondents. It is clear that the training of skills in reading comprehension is not a short-term proposition; Singer and Donlon (1982) corroborate this view with their finding that instruction needs to continue over a variety of material, repeated over multiple trials.

The present study found that there are some items which demonstrated a short-term training effect, although this improvement diminished somewhat across time. A larger sample would be required to provide a more detailed understanding of this short-term effect in relation to ethnicity. The present study suggests that while the effects that were found did not significantly differentiate respondents by ethnicity, the Hispanic treatment group subjects had a slight advantage over their non-Hispanic counterparts, and over the nontreatment control subjects. The Hispanic treatment group gained 1.41 additional correct points from pretest to follow up while the Hispanic contol group lost 1.00 points.

REFERENCES

Arnold, M. Teaching theme, thesis, topic sentences, and clinchers as related concepts. Journal of Reading, 1981, 24, 373-376.

Bateson, G. Steps to an ecology of mind. NY: Ballantine, 1972.

Berk, R.A. (Ed.) Handbook of methods for detecting test bias. Baltimore, Johns Hopkins University Press, 1982.

Bransford, J. D., & Johnson, M. K. Consideraton of some problems of comprehension. In W. Chase (Ed.), Visual Information Processing. NY: Academic Press, 1973.

Bransford, J. D., Nitsch, K. E., & Franks, J. J. Schooling and the facilitation of knowing. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), Schooling and the Acquisition of Knowledge. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Cabello, B. Potential sources of bias in dual language achievement tests. CSE Report, 1981.

California Achievement Test: Technical Bulletin 1, Monterey, CTB/McGraw Hill, 1979.

Dee Lucas, D. & DiVesta, F.J. Learner-generated organizational aids as effects on learning from text. Journal of Educational Psychology, 1980, 72, 304-311.

Doctorow, M.J., Wittrock, M.C. & Marks, C.B. Generative processes in reading comprehension. Journal of Educational Psychology, 1978, 70, 109-118.

Donlon, T., Hicks, M., & Wallmark, M. Six differences in item responses on the Graduate Record Examination. Applied Psych. Measurement, 1980, 4, 9-20.

Gerlen, E. R., & Costar, E. Scoring high in reading: The effectiveness of teaching achievement test-taking behaviors. Elementary School Guidnce and Counseling, 1980, 15, 157-159.

Green, D.R. What does it mean to say a test is biased? Education and Urban Society, Vol. 8, 1975, 35-51.

Harnisch, D.L., & Linn, R.L. Identification of abberant response patterns. Champaign, Illinois: University of Illinois, 1982. National Institue of Education Grant No. G-80-0003, Final Report.

Jensen, A.R. Bias in mental testing. New °York, Free Press, 1980.

Just, M. A., & Carpenter, P. A. (Eds.). Cognitive Processes in Comprehension. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

Kraemer, H. C. Intergroup concordance: Definition and estimation. Biometrika, 1981, 68, 641-646.

Kurata, T, & Sato, T. Similarity of some indices of item response patterns based on an S-P chart. Computer and Communication Systems Research Laboratories, Nippon Electric Company, Research Memorandum E181-4, 1981.

Linden, M. & Wittrock, M.C. The teaching of reading comprehension according to the model of generative learning. Reading Research Quarterly, 1981, 18, 44-57.

McArthur, D.L. Detection of item bias using analyses of response patterns. Paper presented to the Annual meeting of the American Educational Research Association, New York, 1982.

Mielke, P. W., Berry, K. J., Brockwell, P. J., & Williams, J. S. A class of nonparametric tests based on multiresponse permutation procedures. Biometrika, 1981, 68, 720-724.

National Assessment of Educational Progress. Performance of Hispanic students in two national assessments of reading. Denver: Education Commission of the states, 1982.

Petersen, N.S. Bias in the selection rule, bias in the test. In L.J.T. van der Kamp, W.F. Langerak, & D.N.M. deGruiter (Eds.), Psychometrics for educational debates. Chichester, G.B., John Wiley & Sons, 1980.

Rumelhart, D.E.. Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce, and W.F. Brener (Eds.) Perspectives and cognitive psychology, linguistics, and artificial intelligence and education. Hillsdale, N.J. Lawrence Erlbaum Associates, 1980.

Sato, T. A classroom information system for teachers, with focus on the instructional data collection and analysis. Association for Computer Machinery Proceedings, 1974, 199-206.

Sato, T. Analysis of students' pattern of response to individual subtests. Computer and Communications Systems Research

Laboratories, Nippon Electric Company, Research Memorandum E181-2, 1981a.

Sato, T. Similarity of some indices of item response patterns. Computer and Communications Research Laboratories, Nippon Electric Company, Research Memorandum E181-1, 1981b.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Singer, H., & Donlon, D. Active comprehension: Problem solving schema with question generation for comprehension of complex short stories. Reading Research Quarterly, 1982, 17, 166-185.

Sato, T. The construction and interpretation of S-P tables. Tokyo: Meiji Tosho, 1975 (In Japanese).

Sato, T. The S-P chart and the caution index. Nippon Electric Company, Educational Informatics Bulletin, 1980.

Sato, T., & Kurata, M. Basic S-P score table characteristics. NEC Research and Development, 1977, 47, 64-71.

Sato, T., Takeya, M., Kurata, M., Morimoto, Y., & Chimura, H. An instructional data analysis machine with a microprocessor -- SPEEDY. NEC Research and Development, 1981, 61, 55-63.

Scheuneman, J.D. A posteriori analyses of biased items. Unpublished paper, ETS, 1980.

Singer, H. & Donlan, D. Activie Comprehension: Problem solving schema with questin genration for comprehension of complex short stories. Reading Research Quarterly, 1982, XVII, 2, 166-185.

Subkoviak, M.J., Mack, J.S. & Ironson, G.N. Item bias detection procedures: Empirical validation. Paper presented at the Annual Meeting of the American Educatinal Research Association, Los Angeles, 1981.

Ulibarri, D. Cognitive processing theory and culture-loading: A neo-Piagetian approach to test bias. U.C. Berkeley: Unpublished Ph.D. dissertation, 1982.

Ulibarri, D. The test bias issue: A distinction between test bias and test validation. Unpublished paper, 1979.

van der Linden, W.J. A latent trait look at pretest-posttest validation of criterion-referenced test items. Review of Educational Research, 1981, 51, 379-402.

Williams, R.L. Abuses and misuses in testing Black children. Counseling Psychologist, 1971, 2, 62-77.

Wittrock, M.C. A proposal for research in reading comprehension. Unpublished paper, UCLA, 1982.

Wittrock, M.C., Marks, C.B., & Doctorow, M.J. Reading as a generative process. Journal of Educational Psychology, 1975, 67, 484-489.

Table 1

Means and Standard Deviations for
Age, Reading Level, Total Score, and
Subtest Scores by Group and Ethnicity

Group=	Treatment				Control			
Ethnicity= Pre- Post n= Pre- Post- Follow up n=	Hispanic 49 41		Non-Hispanic 17 17		Hispanic 18 18		Non-Hispanic 9 9	
	$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d	$\bar{x}$	s.d.
Age:	9.90	0.55	10.29	0.59	9.87	0.52	10.40	0.55
Reading level	3.51	0.98	3.12	1.87	3.58	1.31	3.91	0.91
Total test score - pre	11.51	5.09	10.65	5.92	13.22	5.74	14.44	5.22
Total test score - post	13.25	4.78	12.88	4.57	13.55	4.56	16.55	4.06
Improvement $\bar{x}$, (%)*	+1.74	(+15%)	+2.23	(+21%)	+0.33	(+02%)	+2.11	(+15%)
Total test score - follow up	12.91	5.28	11.35	5.15	12.22	5.48	15.56	3.91
Improvement $\bar{x}$, %	-.98	(-7%)	-1.53	(-12%)	-1.33	(-10%)	-.99	(-6%)
CTBS subtest score - pre	7.10	3.20	6.76	4.22	9.05	4.19	10.11	2.71
CTBS subtest score - post	8.36	3.30	8.11	3.17	8.72	3.23	10.55	2.07
Improvement $\bar{x}$, (%)	+1.26	(+18%)	+1.35	(20%)	-0.33	(-04%)	+0.44	(+04%)
CTBS subtest score - follow up	7.92	3.42	7.41	3.39	7.83	3.90	9.67	2.60
Improvement $\bar{x}$, %	-.44	(-5%)	-.70	(-9%)	-.90	(-10%)	-.88	(-9%)
Bellagio subtest score - pre	4.41	2.79	3.88	3.05	4.17	3.09	4.33	3.64
Bellagio subtest score - post	4.88	2.27	4.76	2.54	4.83	2.47	6.00	1.93
Improvement $\bar{x}$, (%)	+0.47	(+11%)	+0.88	(+23%)	+.66	(+16%)	+1.65	(+39%)
Bellagio subtest score follow up	5.00	2.44	3.94	2.07	4.39	2.48	5.89	2.26
Improvement $\bar{x}$, (%)	+.12	(+2%)	-.82	(-17%)	-.44	(-9%)	-.11	(-2%)

*"Improvement" values are change relative to immediately preceding test occasion.

Table 2

S-P Analysis by Group by Occasion (Pre-Post)

Group =	Treatment	Control
n =	66	27
S-P D* pre	.658	.566
post	.717	.692
Mean; s.d. C* item pre	.274, .093	.264, .124
post	.303, .115	.312, .167
% items C*>.30 pre	.300	.333
post	.333	.533
Mean; s.d. C* persons pre	.316, .144	.253, .149
post	.339, .128	.297, .121
% person C*>.30 pre	.470	.407
post	.621	.444

Table 3

Proportions of Wrong Responses Above the P-Curve, for CTBS items, by Group by Occasion (Pre-Post)

		Group:	Treatment		Control	
		Occasion:	Pre	Post	Pre	Post
Item		Task				
CTBS (Skyscraper poem)	1*	Recognition and recall	.41	.39	.33	.41
"	2*	Infer main idea#	.35	.22	.31	.20
"	3*	Analysis of structure	.25	.25	.22	.21
"	4	Analysis of style	.50	.40	.46	.31
"	5	Analysis of figurative language	.31	.27	.24	.18
(Abalone story)	6	Translation (rewording)#	.34	.20	.24	.21
"	7*	Literal recall#	.27	.29	.09	.14
"	8*	Literal recall#	.33	.34	.19	.31
"	9	Literal recognition (rewording)#	.29	.31	.24	.13
"	10*	Infer main idea#	.69	.46	.56	.38
"	11*	Relationship (sequence of events)	.59	.40	.75	.75
(Threshing Wheat)	12	Literal recognition#	.57	.46	.45	.83
"	13*	Relationship (literal recall)	.43	.60	.24	.21
"	14	Analysis of structure	.43	.37	.33	.42
"	15*	Translation (rewording)	.40	.52	.50	.44
"	16*	Infer word meaning#	.85	.82	.50	.75
"	17*	Infer main idea#	.57	.57	.21	.08
"	18*	Extended meaning (analysis)	.44	.39	.27	.33
		Mean change		-.04		+.01
		s.d.		.10		.14

* Items with possible bias
# Tasks addressed by treatment intervention

Table 4

Proportions of Wrong Responses Above the P-Curve, for Bellagio Items by Group by Occasion (Pre-Post)

		Group:	Treatment		Control	
		Occasion:	Pre	Post	Pre	Post
Item		Task				
Bellagio (Navaho Chores)	1	Translation (simple rewording)	.29	.37	.33	.60
	2	Infer main idea#	.20	.33	.20	.35
"	3	Infer main idea#	.28	.28	.20	.35
(Legends of the Navaho)	4	Literal recognition#	.35	.40	.25	.27
"	5	Infer main idea#	.35	.28	.38	.57
	6	Inference - extended meaning	.35	.26	.27	.19
(Navaho Beliefs)	7	Relationshp - cause and effect	.73	.63	1.00	.75
"	8	Infer main idea#	.44	.45	.57	.66
"	9	Literal recognition#	.29	.67	.44	.50
(Navaho Shaman)	10	Infer main idea#	.44	.36	.33	.31
"	11	Translation (simple rewording)	.64	.60	.43	.33
"	12	Relastionship (sequence of events)	.78	.81	.56	.83
		Mean change	--		--	
		s.d.	--		--	

#Tasks addressed by treatment intervention

Table 5

Non-Responses by Occasion (Pre-Post)

Pretest	$\bar{x}$ = 4.10	s.d. = 5.83
Posttest	$\bar{x}$ = 0.99	s.d. = 2.41
Difference	$\bar{x}$ = -3.11	s.d. = 5.36

mdn = -3.37
skew = -1.44
range = -24 to +4

Table 6

$\bar{P}$'s and S-P Cautions for Items - N = 40
(CTBS)

	$\bar{P}$				C*		
	pre	post	follow up	"bias"?*	pre	post	follow up
1	.40	.58	.43	yes	.26	.23	.29
2	.40	.58	.48	yes	.21	.11	.25
3	.68	.60	.50	yes	.32	.19	.29
4	.25	.40	.40		.39	.30	.51
5	.60	.63	.48		.24	.19	.19
6	.45	.60	.58		.13	.14	.19
7	.40	.53	.50	yes	.13	.18	.05
8	.43	.50	.53	yes	.17	.27	.23
9	.53	.53	.50		.16	.27	.19
10	.15	.40	.35	yes	.41	.36	.22
11	.25	.20	.28	yes	.33	.33	.36
12	.20	.30	.33		.26	.31	.38
13	.28	.58	.48	yes	.34	.38	.33
14	.40	.43	.58		.19	.28	.29
15	.25	.35	.33	yes	.28	.51	.41
16	.15	.25	.28	yes	.69	.73	.47
17	.28	.30	.10	yes	.34	.25	.31
18	.15	.35	.30	yes	.33	.21	.26
$\bar{X}$	.35	.45	.41		.29	.29	.29
D*	.588	.713	.728				

*Evidence of item bias found in previous study of CTBS (McArthur, 1981).

Table 7

$\bar{P}$'s and S-P Cautions for Items - N = 40
Bellagio

	$\bar{P}$			C*		
	pre	post	follow up	pre	post	follow up
1	.45	.30	.25	.17	.22	.35
2	.65	.60	.55	.28	.28	.17
3	.50	.40	.50	.21	.20	.08
4	.40	.30	.33	.13	.28	.17
5	.43	.58	.48	.26	.31	.28
6	.43	.55	.50	.12	.31	.07
7	.25	.40	.35	.47	.30	.57
8	.25	.20	.20	.18	.19	.14
9	.20	.38	.28	.04	.37	.35
10	.40	.48	.35	.15	.18	.35
11	.23	.23	.25	.23	.30	.10
12	.13	.25	.18	.29	.33	.38
$\bar{X}$	.36	.40	.35	.21	.27	.25
D*	.643	.679	.698			

Table 8

Selected Ratios and Tests of Proportions
(CTBS)

	Moving to Correct* / Moving to Wrong	Always Correct# / Always Wrong	Z pre, post	Z post, follow up@
1	1.11	0.50	1.698 ns	-1.604 ns
2	1.00	1.14	0.894 ns	-1.000 ns
3	0.36	2.00	-0.728 ns	-1.155 ns
4	2.50	0.27	1.732 ns	0.000 ns
5	0.44	2.17	0.243 ns	-1.732 ns
6	2.25	1.18	1.732 ns	-0.348 ns
7	2.00	0.92	1.384 ns	-0.333 ns
8	1.67	1.00	0.728 ns	0.229 ns
9	0.90	1.29	0.000 ns	-0.258 ns
10	3.67	0.06	2.500 p<.01	-0.500 ns
11	1.14	0.06	-0.535 ns	0.832 ns
12	2.00	0.11	1.069 ns	0.302 ns
13	2.60	0.56	2.828 p<.01	-1.000 ns
14	2.17	1.17	0.258 ns	1.342 ns
15	1.50	0.25	1.155 ns	-0.277 ns
16	2.67	0.10	1.155 ns	0.277 ns
17	0.22	0.04	0.333 ns	-2.309 p<.01
18	3.00	0.14	2.530 p<.01	-0.707 ns

* Ratio of number of respondents with permutations 011 or 001 vs number with 100 or 110.

Ratio of number of respondents with permutation 111 vs number with 000.

@ Results of Z-test for correlated proportions.

Table 9

Selected Ratios and Tests of Proportions
Bellagio

	Moving to Correct* / Moving to Wrong	Always Correct# / Always Wrong	Z pre, post	Z post, follow up@
1	0.20	0.38	-0.832 ns	-1.508 ns
2	0.50	2.50	-1.972 p<.05	-0.577 ns
3	1.00	0.91	-1.155 ns	1.155 ns
4	0.75	0.25	-0.943 ns	0.302 ns
5	1.22	0.88	1.500 ns	-1.609 ns
6	1.33	0.86	1.091 ns	-0.577 ns
7	2.00	0.31	1.604 ns	-0.577 ns
8	0.71	0.09	0.378 ns	-0.832 ns
9	1.60	0	1.528 ns	-0.943 ns
10	0.78	0.60	0.728 ns	-1.291 ns
11	1.20	0.05	0.000 ns	0.277 ns
12	1.50	0.04	1.667 ns	-0.832 ns

* Ratio of number of respondents with permutations 011 or 001 vs number with 100 or 110.

Ratio of number of respondents with permutation 111 vs number with 000.

@ Results of Z-test for correlated proportions.

Table 10

Proportions of Respondents with 011 or 001 Permutations

		Treatment Group 011	Treatment Group 001	Control Group 011	Control Group 001
Item: CTBS	1	14	5	11	0
	2	9	9	7	4
	3	9	0	18	7
	4	9	7	0	4
	5	9	3	0	8
	6	16	5	12	0
	7	10	7	4	0
	8	14	14	0	4
	9	12	9	4	0
	10	14	10	18	11
	11	7	16	0	12
	12	16	9	4	12
	13	16	9	8	8
	14	7	17	11	19
	15	5	12	0	0
	16	7	9	8	4
	17	2	5	4	0
	18	9	10	0	4
Bellagio	1	3	10	15	4
	2	12	7	15	19
	3	9	12	4	8
	4	10	14	22	4
	5	17	9	11	4
	6	21	9	11	12
	7	10	9	8	15
	8	7	16	8	15
	9	10	10	7	11
	10	9	19	11	16
	11	3	17	11	8
	12	2	10	0	4

Table 11

Intercorrelations of Eight Pre/Post/Follow up Permutations for Thirty Items

Permutation	111	011	001	010	101	110	100	000
111	1.000							
011	.151	1.000						
001	-.332	-.117	1.000					
010	-.334	.009	.204	1.000				
101	.342	-.053	-.054	.144	1.000			
110	.214	.096	-.392	-.303	.120	1.000		
100	.228	-.172	-.136	-.183	.237	.038	1.000	
000	-.750	-.425	.042	-.014	-.556	-.251	-.393	1.000

Figure 1

Score permutations ordered by item difficulty, CTBS items only. Each band represents the number of treatment participants giving a correct or incorrect response at Pre-testing and Post-testing. The comparable proportions of control participants giving an incorrect Pre-test response but a correct Post-test response are overlaid on the 01 permutation, only.

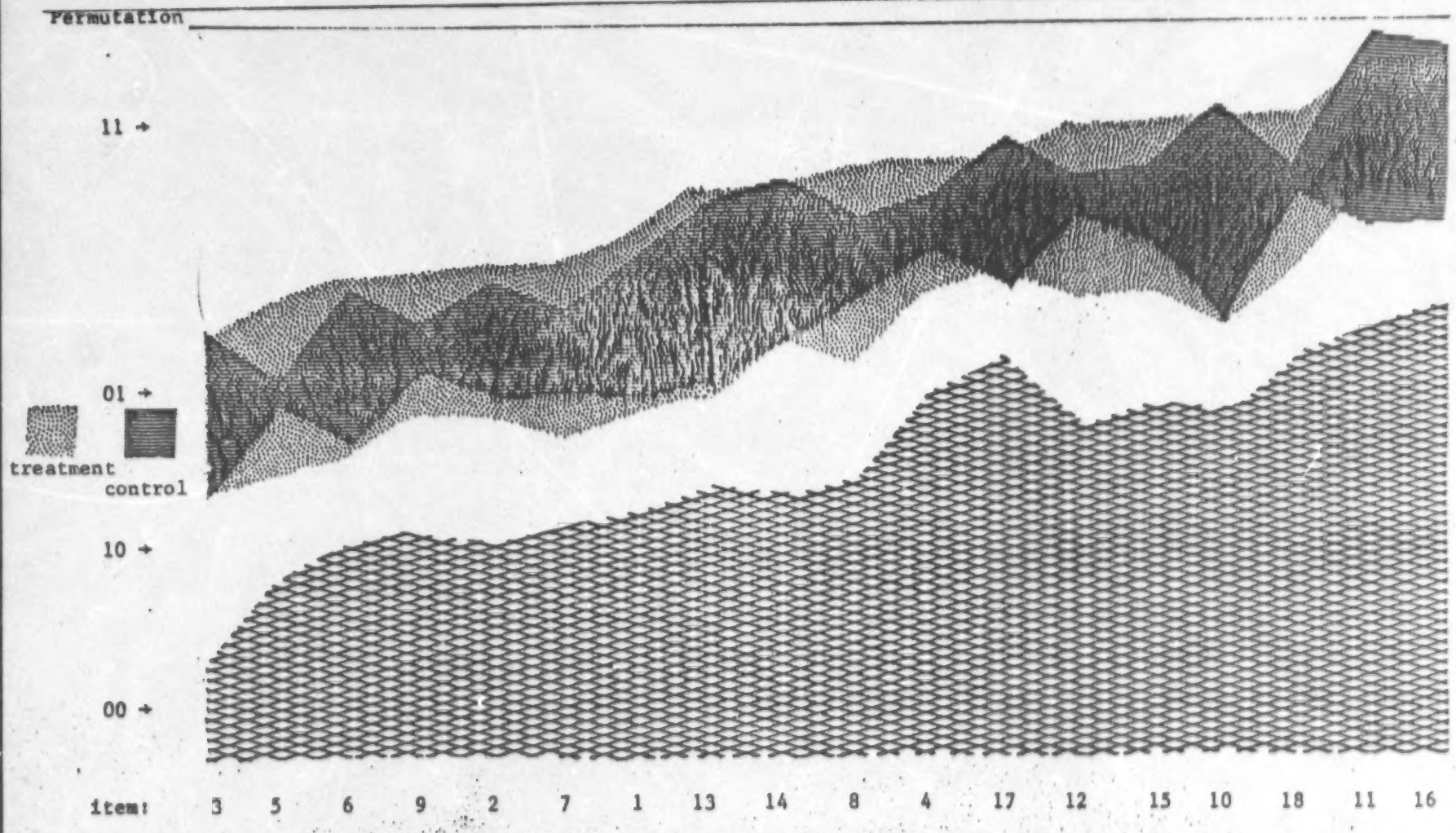
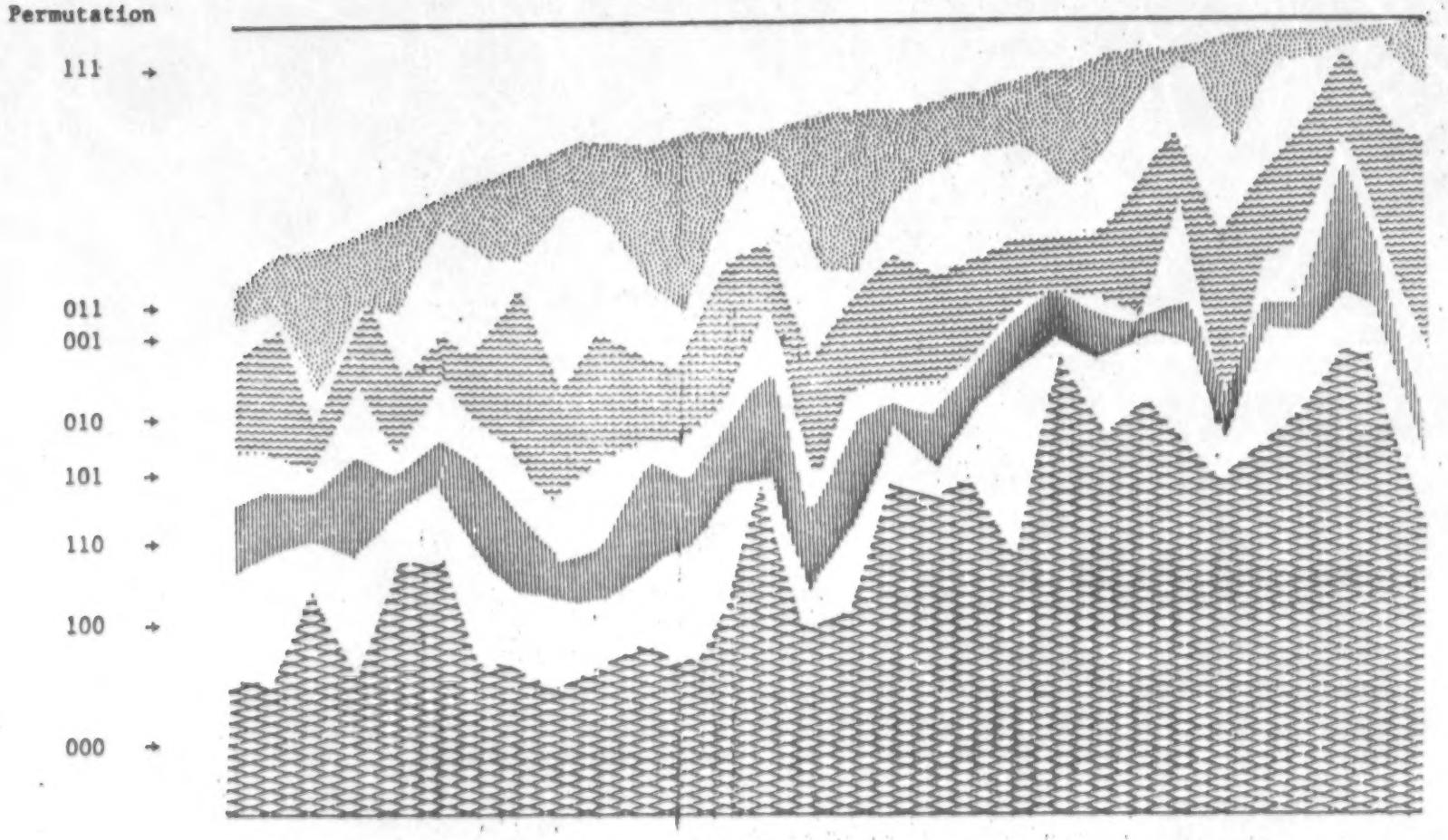


Figure 2

Score permutations ordered by item difficulty. Each band represents the number of treatment participants giving a correct or incorrect response at Pre-testing, Post-testing, and Followup.



B=Bellagio, C=CTBS